Lexical category induction using lexically-specific templates

Richard E. Leibbrandt and David M. W. Powers
Flinders University of South Australia

1. The induction of lexical categories from distributional information

The lexical categories of a language (word classes such as nouns, verbs and adjectives) are of crucial importance in describing its grammar. Several authors (e.g. Maratsos & Chalkley, 1980) have suggested that children might identify the word classes of their first language by using distributional information, i.e. by grouping together words that tend to occur in the same linguistic contexts. A number of researchers have risen to the challenge of producing explicit, computational implementations of this idea. For instance, Redington, Chater and Finch (1998) conducted a corpus analysis in which they obtained typical usage profiles for a number of words, based on the sum of all contexts in which the words had occurred in the corpus. The contexts of a word were taken to be the word that occurred two words before, one word before, one word after and two words after the target word. A cluster analysis was performed to combine words with similar usage profiles into large word clusters which corresponded closely to traditional parts of speech such as nouns and verbs. Similar work by Mintz, Newport and Bever (2002) replicated and extended this result.

One shortcoming of these models is that each particular word type is assigned to only one cluster or lexical category. But in fact, the same word type can be used as a noun, verb, adjective, etc, depending on its usage context (and words are used in this flexible way fairly frequently in the input to children; see for instance Nelson, 1995). The computational models mentioned above can at best identify the *majority* category of a word type, but will make mistakes in categorizing individual instances of words.

It is often the case that a particular context will completely determine the lexical category of a word that occurs in it. For instance, in the sentence frame "Don't X the Y", where X and Y represent slots that may be filled by a variety of words, an adult English speaker knows that when a word appears in the X slot, it is a verb, and when a word appears in the Y slot, it is a noun. A procedure that explicitly lists some of the most important contexts in which words may occur, and assigns a lexical category to a word based on the identity of the *context*, rather than the identity of the word itself, may be expected to deal more effectively with word ambiguity.

A computational approach to lexical categorization that attempts to explicitly identify relevant contexts in this way is the "frequent frames" model of Mintz (2003, 2006a, 2006b). Frequent frames are disjunct contexts consisting of the word immediately preceding a focal word combined with the word immediately following it (i.e. all frequent frames take the form *a X b*, where *a* and *b* are fixed words, and the *X* represents a variable slot). Any word instances occurring in the same frame are categorized together, and frames that have more than 20% overlap in their set of slot fillers are amalgamated into larger, general categories. This technique produces a highly successful categorization of word instances on the basis of their contexts.

There is a large body of evidence to suggest that children are able to categorize words in this way. A celebrated experiment by Brown (1957) showed that 3-year-olds and 4-year-olds are able to make use of nothing more than the linguistic context in which a novel word occurs in order to guess at its meaning. Children were exposed to a target picture of, for example, a pair of hands performing an unusual kneading motion on an unfamiliar substance in an oddly-shaped container. Three additional test pictures contained only one of the components of the original picture (the motion, the substance or the container). Children were introduced to a novel word (say, "sib") in one of three linguistic conditions: mass noun ("here you can see *some sib*"), common noun ("here you can see *a sib*"), or verb ("here you can see *sibbing*"), and when asked to pick out another instance of "some sib", "a sib" or "sibbing" from the test picture set, reliably chose the unusual substance, the container, or the kneading motion respectively.

One conclusion that can be drawn from this experiment is that the lexical category assigned to a word does not necessarily depend on the various contexts in which that word has previously been used in the child's experience (because the children had not heard the novel words used in this experiment before). Brown's experiment shows that, at least at the age of three or four, children are able to categorize an unknown word after a single exposure, based solely on the context in which the word was used, and to make a semantic interpretation of the word based on that categorization.

Mintz (2006a) has provided evidence suggesting that infants as young as 12 months of age may be able to use the distribution of a word in frequent frames to categorize that word (even in the absence of any visual information to which the word could be "anchored"). Infants were exposed to four novel words, two of them used in noun frames and two in verb frames. Most (but not all) of the frames used were frequent frames. In a test using the preferential head-turn procedure, infants listened longer to sentences that seemed to be incompatible with their initial experience than to sentences that were consistent with it. For instance, if a nonsense word had been introduced in a verb frame during familiarization, infants listened longer when they heard the same word embedded in a noun frame at test. This result suggests that even 12-month-old children may be able to distinguish between a number of English noun frames and a number of verb frames, and hence that the frames themselves have some psychological reality for children even at this early age.

However, it should be noted that some researchers have concluded that children do not command adult lexical categories until a much later age. For instance, Olguin and Tomasello (1993) have shown that 25-month-old children are reluctant to use a novel verb in contexts other than the one in which they have heard it modeled. By contrast, 23-month-olds easily extend the use of novel nouns to a variety of contexts (Tomasello & Olguin, 1993).

The contexts considered in the frequent frames approach are restricted to a very specific *a X b* structure. It seems that this structure fails to cover many stereotypical frame patterns in English that we might intuitively believe to impose a lexical category on the words that occur in them, e.g. the question "do you X?", the imperative utterance "X it", the noun phrase "the X" or the part-verb phrase "going to X". In this paper, we attempt to extend the work of Mintz (2003) to accommodate contexts similar to these examples. Frequent frames seem to represent a "topological" approach to defining the context of a word, i.e. in terms of words that occur in fixed relative positions to a focal word. In considering alternative ways of defining a word's context, one possibility is to search for some of the most common *constructions* in English.

2. Constructions and lexically-specific frames

From the point of view of the linguistic theories that go under the banners of Cognitive Grammar (e.g. Langacker, 1987) or Construction Grammar (e.g. Goldberg, 1995), the units of a language are constructions, form-meaning pairs that exist at various levels of granularity (e.g. morphemes, words, phrases, clauses). Under this approach, a particular utterance may be represented in a speaker's or listener's language system at various simultaneous levels of abstractness. In particular, certain constructions are made up of a sequence of specific words, combined with a number of slots that are filled by variable material. For instance, the slots in "the X the Y" may be expanded to produce the construction "the sooner the better".

Some language development researchers investigating children's syntactic development have suggested that such *lexically-specific frames* (also called item-based or mixed constructions by Tomasello, 2003) may play a prominent role in children's early linguistic knowledge. Lieven, Pine and Baldwin (1997; see also Pine and Lieven, 1993), based on an analysis of speech data collected from a group of children between their first and third birthdays, have suggested that many of these children's productions can be accounted for by the use of a relatively small set of semi-schematic utterance patterns. Some examples of these patterns are "it's a X", "me got X", "want X", "oh don't X" (where the X's represent the variable slots). The usefulness of these item-based frames lies therein that they provide a way for the child to move from verbatim, memorized utterances to adult syntactic competence by way of an intermediate position where only part of a construction is abstract, while the remainder is grounded in concrete items. Lieven et al. (1997) further suggest that lexically-specific frames may provide a route by which lexical (and other grammatical) categories are learned.

While these analyses have focused mainly on productions *by* children, Cameron-Faulkner, Lieven and Tomasello (2003) have analysed a corpus of mothers' speech *to* their children in order to identify the most frequently-used constructions, and found that a set of only 52 item-based constructions (e.g. "That's …", "Shall we …", "What's …") was sufficient to account for over half of mothers' utterances in the corpus, and moreover that children's use of the most common of these constructions correlated highly with that of their mothers.

These corpus-based studies show that a great number of lexically-specific constructions in English are very common in the input to children, and in children's own speech. If it is true that the child needs to master the constructions of a language in order to attain adult syntactic competence, then inducing lexical

categories from item-based constructions is an attractive idea, as it makes it possible to provide a unified account of both syntax learning and lexical category induction.

In this work, we attempt to reconcile Mintz's (2003) context-centered approach to word category induction with the work by Lieven et al. (1997) and Cameron-Faulkner et al. (2003) on the importance of lexically-specific frames in language learning, by making use of contexts that are likely to be lexically-based constructional frames.

In the first and second of the three experiments in this paper, we present a procedure (implemented in a computational model) to identify a number of lexically-specific frames / constructions in a corpus of child-directed speech, and demonstrate that the frames that are discovered are adequate for the induction of the three major content-word lexical categories (nouns, verbs and adjectives).

Experiment 1 focuses on finding schematic structures for complete utterances. Many utterances to children are either valid statements, questions, or requests, or else utterance fragments that are often constituents such as noun phrases (see Cameron-Faulkner et al., 2003), and so complete utterances may be regarded as examples of the most basic constructions in the input to children. Tomasello (2003, p. 113ff.) argues that *utterance-level constructions* play a prominent role in language development: these are verbal expressions that can be used as complete utterances, and that are associated in a routinized way with certain communicative functions. While we do not make use here of information about meaning or communicative intent, and so are not identifying utterance-level constructions directly, we nevertheless attempt to discover some of the most prominent full-utterance structures in the corpus from textual and distributional information alone. In Experiment 2, we extend this model by using a substitution test to identify hypothetical (typically phrase-like) linguistic constituents that occur nested in larger utterances.

3. Experiment 1: A procedure to discover full-utterance templates

In English, there seems to be a prominent role for *function words* in constituting the lexically-specific portions of item-based frames. For instance, many of the structuring elements in the frames identified by Lieven et al (1997) and also by Cameron-Faulkner et al. (2003) were function words such as "the", "me", "don't", etc. Work by Shady (1996) has shown that children at the age of 16 months (but not yet at 12.5 months) are sensitive to violations in the co-occurrence patterns of function words in utterances; for instance, they listen longer to a passage containing correctly-formed sentences such as "*the* large cake *is* baking" than to one containing modified, ungrammatical sentences such as "*is* large cake *the* baking". Shady (1996) also found that 10.5-month-old infants preferred listening to correctly-formed sentences rather than to ones in which all function words were replaced with nonsense words, but that, surprisingly, no listening preference either way was exhibited when the function words were retained and several of the *content* words were replaced with nonsense words instead.

The function words seem to bear a great deal of the brunt in providing structure to utterances in English, and it is tempting to consider that there might be a basic dichotomy at work in English between function and content words, such that sentence structures might be described using function words alone, with slots in the positions where the content words should go. However, in implementing a procedure to discover these frames, we cannot merely identify the function words from our own knowledge of English; the language-learning child cannot necessarily be assumed to know from the outset which words are functional and which contentful.[1]

Furthermore, Tomasello (1992, 2003) provides evidence to suggest that many *verbs* are the organizing elements around which constructional patterns are formed in children's early productions, so that function words may not be the only relevant candidates to be considered in constructing item-based frames.

Work by Gómez and Maye (2005) on artificial language learning may shed some light on the process of learning item-based frames. Fifteen-month-olds, but not 12-month-olds, were able to learn to identify valid sentences from an artificial language in which sentences conformed to an *a X b* structure, with the X slot representing a variable element. Gómez and Maye also found that learning was facilitated by increasing the number of word types that appear in the X slot during training.

[1] There certainly are a number of phonological cues that can be used to distinguish English function words from content words (see Morgan, Shi and Allopenna, 1996; Shi, Werker and Morgan, 1999). Nevertheless, we are interested here in finding techniques that can identify templates from corpora that are not annotated with phonological information, and so we will not consider phonological issues here.

If we extrapolate the results of Gómez and Maye (2005) to natural language learning, it might be the case that an important prerequisite for learning about discontinuous frames is that there should be a large number of different filler types appearing in the slot of each frame. If this interpretation is correct, a frame needs to be attested several times in the corpus, in the form of utterances that conform to the frame but have different slot fillers in each case. Note that, if we were to count the *frequency of occurrence* of words in the corpus, the lexically-specific words would receive a greater contribution to their total from their occurrence in the template than the slot fillers, as they appear every time the template appears, whereas the slot fillers do not. If we add to this the assumption that there is only a restricted set of words that are likely to be used as the lexically-specific part of a template, then it is only these words that will ever enjoy this frequency advantage. Hence, it follows that the words that are the structuring elements in lexically-specific frames are probably to be found in the set of the most *frequently-used* words in a corpus.

Our template discovery procedure is therefore the following: First, find the set of the most commonly-occurring *N* words in the corpus. (In these experiments, *N* is set at 150 throughout; manipulation of the value of *N* affected the set of templates that were discovered, but did not greatly affect the quantitative evaluation of the lexical category assignment process that we will consider later.) Next, rewrite every utterance in the corpus, retaining each word that occurs in the list of the top 150 most frequent words in the corpus, and replacing every other word with an X. Treat each rewritten utterance as a potential template, and each X as a potential slot in the template. Collect all templates that have occurred in the corpus with at least 5 different words occurring in their X slots. (At the same time, these filler words are required to occur in at least 5 different templates.) The templates that remain after these constraints are applied comprise the set of lexically-specific templates produced by the procedure. Any words in the set of 150 that were not taken up into templates are "returned to the pool" of slot-filling words, and are replaced with X's.

It is quite plausible that the process by which a language-learning child discovers the lexically-specific templates of her native language might follow a similar route. In the course of being exposed to language input, it can be expected that the child will initially recognise no words, and at later stages will be able to recognise an increasing number of words. It seems plausible that the first words she will be able to recognise from their phonological strings alone will be the most *frequent* words. Furthermore, if the child is able to notice co-occurrence patterns between words in an utterance, she will, once again, most likely start with the co-occurrence patterns between the most frequent words. Suppose that at some stage the child can recognise the very familiar words "you", "can't" and "that", but not yet the less frequent word "chew". When faced with the utterance "you can't chew that", what the child can recognise out of the utterance could be represented as "[you] [can't] […] [that]". Given more extensive experience of this pattern, possibly with different slot fillers ("eat", "drink", "have", etc.), the child may eventually discover the co-occurrence pattern between the words, so that the larger pattern "you can't … that" may become a familiar one. Note that we are referring here to recognition only, and to a process by which the "texture" of English utterances becomes familiar to the language-learning child; it is *not* required that the child should know what any of these structure-building words mean. Compare this situation with the one facing Shady's (1996) infant subjects, who seemed content to listen to sentences that preserved the function-word co-occurrence patterns of English, even if the content words occurring between the function words were nonsense.

We implemented this procedure on the Manchester corpus (Theakston, Lieven, Pine & Rowland, 2001) from the CHILDES database (MacWhinney, 2000). We made use only of the speech by mothers to their children, and pooled the data from all 12 mothers. Out of the 150 most frequently-occurring words, only 136 words productively combined with other words in order to form templates; these words are shown in Table 1 in descending order of frequency. While most of these words were function words, there are a number of verbs such as "want", "see", "come", "make", etc. and one rather concrete noun ("car"). The corpus was rewritten so as to retain only these words; this left us with a final set of 1240 templates[2], with 1113 different words occurring in their slots.

[2] For the purpose of this paper, we applied two additional heuristics to constrain the set of templates discovered by the computational procedure, both of which have been found in prior testing to improve the quality of the final lexical category assignment. The first is that no templates contain a consecutive sequence of X's. The second is that all templates start with specific words, and not with X.

Next, frequency counts were collected of the number of times that particular words occurred in particular templates, resulting in a co-occurrence data matrix where rows represent templates[3] and columns represent words. The rows of this matrix (the templates) were subjected to a hierarchical clustering analysis, using a distance measure based on the Spearman rank correlation, and the average link clustering algorithm of Sokal and Sneath (1963). Hierarchical clustering produces a tree structure, which can be "cut off" at different levels to produce different numbers of mutually exclusive clusters. In these experiments, we choose to cut the tree so as to produce three clusters of templates. Larger numbers of clusters still produce intuitive results, but we are interested in trying to obtain a clustering corresponding to the three "main" lexical categories: nouns, verbs and adjectives.

The assignment of lexical categories to individual word instances is now straightforward; a particular word in an utterance is assigned to lexical category *K* if and only if the template in which it occurs belongs to cluster *K*. (Note that, of course, the clustering algorithm does not have knowledge about the labels noun, verb or adjective, and hence cannot use these labels.) This means that a word is assigned to a category based on its context alone, regardless of which word it actually is.

you, the, it, a, to, oh, that, what, is, on, I, do, and, in, there, are, we, that's, no, one, your, it's, have, [the child's name], don't, can, right, he, going, well, this, not, go, got, put, then, look, want, yeah, now, think, of, what's, with, like, for, they, all, did, you're, yes, here, get, isn't, me, see, come, them, some, she, shall, up, out, be, okay, just, mmhm, at, mummy, was, know, there's, her, he's, very, good, you've, where, bit, little, didn't, because, down, gonna, off, does, doing, big, back, him, I'm, can't, his, make, about, where's, they're, why, doesn't, more, say, my, play, again, nice, these, over, but, car, who's, thank, aren't, has, what're, two, let's, baby, who, other, those, daddy, or, another, haven't, I'll, how, take, gone, she's, need, please, were, find, any, away, too

Table 1. The 136 words from the Manchester corpus used to form lexically-specific templates.

Some representative templates from each of the three template clusters are shown in Table 2. In templates with more than one slot, the active slot is indicated by an X, and the other slots by Z's. Note that the members of Cluster 1 all seem to be templates that can readily accept verbs into their slots. Likewise, templates from Cluster 2 and Cluster 3 seem to be amenable to, respectively, adjectives and nouns (although there are some errors; note "I don't know X" in Cluster 2, or "that was a big X" in Cluster 1).

CLUSTER 1	CLUSTER 2	CLUSTER 3
are you going to X her ?	a X car ?	a X
can you X a Z ?	a X one ?	and your X
can you X it ?	a bit X , isn't it ?	did you like the X ?
did it X ?	are we X ?	have you Z a X ?
do you want me to X it ?	he's X ?	here are the X
don't X it	I don't know X	in his X
I'll X the Z	I know it's X	let's make a X
it doesn't X	is it X ?	more X
let's X	it was X	no X ?
mummy X it ?	it's not X	on Z of the X
oh that X	make it X	put your X on
shall we X it ?	she's X , isn't she ?	some X ?
that was a big X	that one is X	that's not your X
what Z did you X ?	then the X	the baby X ?
what're we going to X ?	very X	this is your X
where do you X ?	what Z to X ?	what X have you got ?
you X your Z	what's X ?	what about these X ?
you can't X it	what's he X ?	what does a X say ?
you have to X it Z	you are X	where's my X ?
you're going to X	you're all X	your X

Table 2. Representative templates from each of the three full-utterance template clusters.

[3] Note that many templates contain more than one X slot. Strictly, we are speaking here of template *slots* rather than templates, so that two different slots in the same template would be represented as two different rows in the data matrix. To avoid clumsiness in exposition, though, we will use the term 'templates' throughout.

It may be easier to understand the nature of these clusters when they are represented in terms of the words that most commonly occur in the member templates. Table 3 lists the 40 most "prevalent" words occurring in the templates of each cluster. These word lists were created by counting the number of distinct templates out of each cluster in which a particular word appeared in the corpus, then sorting the words according to their counts, so that the words which appeared in the greatest number of different templates from, say, Cluster 1, appeared at the top of the word list for Cluster 1. Of course, these lists are of words appearing in context, and so it is perfectly possible for the same word to appear on more than one list: note for instance that "drink" appears on the list for both Cluster 1 and Cluster 3, corresponding to its usage as a verb and as a noun respectively. These lists make clear the strong verb-like, adjective-like and noun-like characters of the three clusters. The only clear "anomalies" occur in the adjective-like Cluster 2, with the presence of a number of names, e.g. "dolly", "Gordon" and "Thomas", as well as a number of present participial forms (which were nevertheless used to describe states of various protagonists, and hence were used in an arguably adjective-like way).

In order to evaluate these clusters quantitatively, a procedure is followed which is becoming standard practice in this field. The lexical category assignments made to word instances in the above scheme are compared against a correct classification. The compilers of the Manchester corpus have manually assigned part-of-speech tags to all the words in the corpus; this assignment was used as the correct "gold standard". Comparing against the gold standard, the numbers of *true positives*, *false positives* and *false negatives* are calculated, abbreviated respectively as TP, FP and FN. A true positive is registered whenever two words are assigned to the same category in the correct classification, and also in the empirical classification obtained from the template clusters. A false positive is registered when two words are assigned to the same category by the empirical classification, but actually belong to two different categories according to the correct classification. A false negative is registered when two words that belong to the same category according to the correct classification are assigned to different categories by the empirical classification. The quantitative measures used to express the degree of success of a categorization are based on these three numbers, and are known as *accuracy* and *completeness*. Accuracy is defined as TP / (TP + FP), and represents the proportion of word pairs put together by the empirical classification which belong together according to the correct classification. Completeness is defined as TP / (TP + FN), and expresses which proportion of word pairs that belong together according to the correct classification are actually put together by the empirical classification. There is typically a trade-off between these two measures, and it is customary to summarize them in a single measure, namely the harmonic mean of accuracy and completeness, known as the *F* value.

CLUSTER 1	CLUSTER 2	CLUSTER 3
eat, sing, open, read, draw, hold, move, build, fix, hurt, catch, count, hear, help, pull, remember, try, drive, use, break, drink, turn, hide, bite, blow, push, tell, reach, close, fit, forget, kick, bang, choose, cook, crash, fall, fetch, finish, jump	red, broken, stuck, green, blue, naughty, yellow, alright, Thomas, tired, yours, hiding, hot, done, cold, eating, pink, crying, funny, lovely, better, coming, dirty, hard, poorly, sleeping, silly, dolly, hungry, finished, wet, clever, Gordon, lost, playing, purple, happy, heavy, orange, sad, white	train, horse, cow, man, bridge, house, pig, box, cake, dog, tiger, hat, book, fish, cat, boat, monkey, door, drink, eggs, sheep, tower, chicken, foot, ball, penguin, elephant, head, water, bag, bricks, nose, duck, animals, picture, truck, giraffe, table, tractor, hand

Table 3. The 40 most prevalent words (see text) occurring in templates from each of the three full-utterance template clusters.

The results of the categorization comparison just described are shown in the first column of Table 5 (labeled "Full-utterance templates (One-dimensional)"). The categorization obtained by assigning a word to the category into which its template has been clustered proves to be fairly correct, attaining an F score of 0.748, as compared against the value of 0.434 that would have been obtained had words been assigned to categories at random.

4. Experiment 2: Extending the discovery procedure to nested templates

While the procedure outlined above was fairly successful in inducing the three main lexical categories from full-utterance templates, it must be noted that there is a great deal of redundancy in the templates that that procedure finds. Templates such as "Find the X", "Are you Z the X ?", "Are you going to Z the X ?", "Can I have the X ?", etc., are all assigned to the "noun" cluster; yet we might suspect that it is just the local noun phrase structure "the X" that is doing the work in these cases of identifying the word in the X slot as a noun. Furthermore, one could surmise that the prevalence of "the X" in the above contexts and many others is due to the fact that it is a linguistic *constituent*, i.e. it is a coherent unit which can be embedded in a variety of contexts. It would be of great use, in learning about lexical categories, to be able to identify these nested constituents. If the phrase "the X" was identified as a nested constituent in all of the above larger templates, then the templates themselves could be discarded in favour of "the X" only, and their word occurrence data, which had been divided among a set of independent templates, could now be credited to the single "the X" template, thereby making the clustering process more compact and accurate.

A traditional test for a linguistic constituent holds that multi-word constituents in an utterance can often be replaced by a single word. This test forms the basis for the procedure outlined here, which will attempt to identify regularly-occurring smaller templates embedded in full utterances.

Suppose that a child hears the utterance "Do you want grapes ?" and some time later, "Do you want some grapes?" The first utterance would be represented schematically under our approach as "Do you want X ?" and the second as "Do you want some X ?". In the first utterance, any word that goes into the X slot is by assumption a linguistic constituent (because we have taken the words of English as our starting point and have assumed that there is a way to segment utterances into words). Now it is possible that the juxtaposition of the second structure against the first suggests to the child the possibility of extending the set of slot fillers in "Do you want X ?" to include also the multiword structure "some X". Supporting evidence for this hypothesis would accrue if it could be shown that other multi-word structures can also appear in the slot of "Do you want X?", and if those multi-word structures can be shown to be embedded also in a variety of other template slots (thereby confirming their constituent nature).

The process of discovering nested templates is as follows: All pairs of utterances in the corpus, rewritten so as to replace less-frequent words with X's[4], are compared against each other. Utterance U_1 is *schematic for* utterance U_2 if U_2 can be transformed into U_1 by substituting some sequence of words in U_2 by an X. In processing the entire corpus, it is possible for extended chains of schematicity to be discovered. For instance, given the four schematic utterance structures "do you X", "do you want X", "do you want me to X", and "do you want me to get your X", the algorithm describes each structure in the above sequence as schematic for the one after it, as each structure can be elaborated into the one that follows it in the chain by replacing an X with a multi-word sequence that is hypothesized to be a possible linguistic constituent. If this hypothesis is correct, one would expect to see other utterance pairs where one utterance is elaborated into the other by replacing an X with the same multi-word sequence.

This process of matching sentences against each other is exactly the one used by Van Zaanen (2001) in his work on Alignment-Based Learning (ABL), a computational technique aimed at automatically discovering syntactic structure in a corpus. The only differences are (i) that we start from a corpus that has already been redescribed in terms of frequent words and X's, and (ii) that we consider only alignments where a single X is replaced by a sequence of words, whereas Van Zaanen considered all possible transformations between utterances. Both of these constraints drastically reduce the number of possible hypotheses that are considered.

A schematicity chain provides a kind of structural bracketing for the last utterance in the chain; the bracketing can be constructed by placing each putative constituent in a pair of brackets, potentially producing a multi-level hierarchically nested structure. For instance, the chain above is represented by the algorithm as "do you [want [me to [get your X]]] ?" From a bracketed structure, we can collect co-occurrence data, just as we did in the previous experiment for words that occurred in full-utterance template slots. When a slot can be filled with a filler consisting of more than one word, the slot will be indicated using a Y instead of an X. The example structure shows us that "do you Y" can be filled by "want Y", "want Y" can be filled by "me to Y", and "me to Y" can be filled by "get your X". Once all of the nesting template – nested template co-occurrence data has been collected, we can discard unreliable data (as

[4] Again with the proviso that utterances with sequences of consecutive X's are not considered.

before), by dropping from the data matrix all nesting templates that have fewer than 5 different structures appearing in their Y slots, and all nested structures that occur in fewer than 5 nesting templates. Some examples of nested templates and their associated nesting structures are shown in Table 4.

It would at this point be possible to cluster the co-occurrence data, potentially allowing a higher-order *syntactic* categorization where multi-word fillers are assigned to categories according to the larger contextual structures in which they are nested. We do not explore this idea here, but return to the issue of lexical categorization; bear in mind that we are still interested in template-word co-occurrence, but simply wish to find nested templates in addition to full-utterance templates.. In the bracketed structure above, the most deeply-nested putative constituent containing the X slot from the original utterance "do you want me to get your X ?" is "get your X". If "get your X" is indeed a constituent, then it is the most immediate context that is relevant to the categorization of the word that fills the X slot. Therefore, we can use this information in order to parse the corpus again, collecting template-word co-occurrence data as before for the purpose of lexical categorization. Now when the algorithm encounters one of the original utterances from which the above example was derived, e.g. "do you want me to get your rolling-pin ?", this is stored in the data matrix as an instance where "rolling-pin" occurs in "get your X", rather than in "do you want me to get your X ?" as before. In this way, then, we parse the corpus once again for template-word co-occurrences, this time choosing the most deeply-nested template we can find as the template context in which a word occurs. In cases where no nested template can be found, the algorithm "falls back" to the full-utterance template as before. This potentially provides a more accurate representation of the data, and also allows a larger number of utterances to be used than was the case with full-utterance templates alone.

your X:	***very X***:	***going to X***:
about [your X]	are they [very X]?	are you [going to X]?
at [your X]	go [very X]	he's [going to X] is he?
do [your X]	he's [very X]	it's [going to X]
I'm [your X]	is it [very X]?	like [going to X]
in [your X]	look [very X]	not [going to X]
put [your X] on	not [very X]	she's [going to X]
there's [your X]	that's [very X]	we're [going to X]
who's [your X] ?	you're [very X]	who's [going to X]?

Table 4. Selected nested templates and the immediately surrounding contexts in which they occur.

There still remains the issue of how we should go about parsing the corpus with the newly-discovered nested templates. For example, one of the discovered nested templates is the "archetypal" noun phrase structure "the X". If the parsing algorithm recognizes this template in a *context-free* manner, i.e. everywhere that it occurs regardless of the surrounding context, then the co-occurrence data that it collects will cover instances where the filler of this template is a noun, e.g. "the tower", "the giraffe", "the eggs", etc., but also instances where the filler is an adjective, e.g. "the red" when it occurs as part of a longer utterance "the red door". This "muddy" information will necessarily confound the clustering process.

The solution taken in the experiment reported here is to recognise a nested template only in those contexts where there already exists evidence to suggest that it is acting as a *whole* constituent. Such a situation occurs when the template is nested inside one of the contexts in which it was initially discovered during the alignment process.

The entire corpus is therefore parsed again in order to find nested templates embedded inside nesting templates, and occurrence data is collected describing which words occur in the nested templates. This produces a matrix of 656 templates by 1465 focal words.

As before, the set of nested templates (and remaining full-utterance templates) is subjected to clustering according to the profiles of words that occur inside them. Categorization is done by assigning each word instance to the category corresponding to the cluster of the template in which it occurs, and is evaluated against the "gold standard" categorization as before. Results are shown in the second column of Table 5. The extended set of templates was slightly more successful in categorization than the set of full-utterance templates, attaining an F score of 0.809.

5. Category induction from ambiguous words and contexts

The previous two experiments were reasonably successful in classifying words as belonging to the three main lexical categories solely from their positions in some of the most common full-utterance and nested frames in English. However, some mistakes in categorization did occur, as shown in the evaluation measures in Table 5. Part of the problem lies in the fact that some templates produced by the two discovery processes can legitimately accept words from more than one lexical category in their slots (something which is largely not the case with frequent frames, for instance). Some templates such as "Are you going to X ?" can accept either a verb ("play") or a noun ("playgroup"), and thus are partially informative, but ambiguous (there are really two constructions here, organized around two different meanings of "going to"). Other templates such as "and X" or "oh X" contain lexically-specific words which are not very closely associated with the words in the slots, and hence are almost entirely uninformative.

In a similar vein, Pinker (1979) has criticized distributional theories of lexical category induction by asserting that the task is intractable, *inter alia* because of the large amount of ambiguity prevalent in everyday language. Given the sentences "Jane eats turkey", "Jane eats fish" and "Jane can fish", Pinker suggests that a child following a distributional strategy might erroneously accept "Jane can turkey" as a valid sentence, due to the ambiguity of the word "fish", which acts as a noun in one sentence and a verb in another. By the same token, contexts can also be ambiguous; a distributional analysis that starts from "Jane eats turkey", "Jane eats slowly" and "The turkey is good" would supposedly accept "The slowly is good" as a valid sentence, because the frame "Jane eats X" does not uniquely pinpoint the category of the word occupying the X slot. Although Pinker's analysis assumes a rather primitive form of distributional analysis (see Redington et al., 1998, for a critique), it is nevertheless true that models that assign all instances of a particular word to the same category (e.g. Redington et al. 1998), or assign all instances of words in a particular context to the same category (e.g. Mintz, 2003), will be prone to the kinds of errors that Pinker identifies.

Note, however, that in the "Jane can turkey" example, there is likely to be a great deal of distributional information from other utterances to suggest that "Jane can X" is a frame that favours verbs only, whereas "turkey" is nearly always used as a noun. *Combining* these two sources of information might be enough in itself to resist the generalization to "Jane can turkey", as the context and word combination would be in conflict. Furthermore, hearing "fish" appear in the same context ("Jane eats X") as the reliable noun "turkey", and subsequently in the reliable verb context "Jane can X", could prompt the child to explicitly flag the word "fish" as ambiguous, and therefore an unreliable basis for categorial generalization. In this way, the child could avoid extrapolating from "Jane can fish" to "Jane can turkey".

These considerations suggest that combining category information from both the word and the context in which it occurs may provide for a more accurate categorization strategy than taking only one of these two sources of information into account. An important insight from studies such as Redington et al. (1998) or Mintz et al. (2002) is that for most word types in the input to children, there is a "majority" category to which each word type most often belongs, with many words only ever being used in one category. In Experiment 3, therefore, we turn our attention to obtaining an improved categorization that combines template and word categorization information.

6. Experiment 3: Combining information from words and templates

In devising a strategy for combining word and template clustering information, we cannot merely cluster templates into categories to obtain one categorization, cluster words together to obtain another categorization, and then combine the two categorizations, because, in general, there is no way to determine which categories from the one clustering map onto categories from the other clustering. What is required instead is a way to *simultaneously* cluster words and templates to the same set of categories, i.e. a kind of "co-clustering" approach.

While many sophisticated co-clustering algorithms have been developed, particularly for analysing gene expression data (see Madeira & Oliveira, 2004, for a review), a relatively simple approach is followed here. Starting with the initial set of template clusters, we attempt to express, for every word, the probability that it is associated with each particular cluster, and then do the same for every template. (As initially every template is allocated to only one cluster with a probability of 1, this step entails making the initial clustering "fuzzier".) To express the probability that word w_j is associated with a particular cluster c_k we

simply count the number of different templates in which w_j occurs, and ask what proportion of these come from cluster c_k (according to the original clustering). Expressed as a formula, we have according to Bayes' rule that

$$P(c_k \mid w_j) = \frac{P(c_k, w_j)}{P(w_j)} = \frac{\sum_{i : t_i \in c_k} b(d_{ij})}{\sum_i b(d_{ij})},$$

where d_{ij} gives the number of occurrences of word w_j in template t_i and $b(d_{ij})$ is equal to 1 if $d_{ij} > 0$ (i.e. if word w_j occurred in template t_i), and is equal to 0 if $d_{ij} = 0$.

The probability that a template t_i is associated with a particular cluster c_k can now be calculated by considering the set of all words that occur in the template, and adding together the probabilities $P(c_k \mid w_j)$ that each of these words is associated with c_k (as calculated before), then dividing by the total number of words occurring in the template. Bayes' rule gives

$$P(c_k \mid t_i) = \frac{P(c_k, t_i)}{P(t_i)} = \frac{\sum_j P(c_k \mid w_j) b(d_{ij})}{\sum_j b(d_{ij})}.$$

For each word and each template, we now have a "category profile" specifying the probability that each word or template is associated with ("belongs to") each cluster. The only remaining detail concerns how the word and template information are combined in order to assign a lexical category to a particular word occurring in a particular template. The solution, for each word w_j occurring in template t_i, is simply to *add* together $P(c_k \mid w_j)$ and $P(c_k \mid t_i)$ for each of the clusters c_k, then pick the cluster (category) with the highest probability sum as the category to which the word instance belongs.

As an illustration, consider the following two examples taken from the simulation of this algorithm, shown in Figure 1 and Figure 2. The word "mean" is ambiguous in its lexical category, and can be used as either a verb or an adjective (and occasionally as a noun, but probably not often in the input to children). This is reflected in the categorial probability vector of "mean": the probability for Verb is 0.51, and for Adjective 0.45. When used in template contexts that are fairly unambiguous, however, the combination of both word and template information makes it possible to disambiguate between the two kinds of usage of "mean". The template "What do you X?" is biased towards verbs, and the template "That's X" is biased towards adjectives. When the probabilities from words and templates are added together, these template biases are sufficient to tip the balance; the result is that "mean" in "What do you mean?" is categorized as a verb, and "mean" in "That's mean" is categorized as an adjective. In similar fashion, the template "Not X" is somewhat ambiguous, though biased towards adjectives; when the word type embedded in "Not X" is also biased towards adjectives, as is the case for "cold", then the categorization chosen is Adjective; but when the word "beans", which is strongly biased towards nouns, occurs in the template, the word type bias dominates and the word instance is categorized as a noun.

	What do you X?		mean		What do you mean?
N	0.03		0.04		0.07
V	0.70	+	0.51	=	**1.21**
A	0.27		0.45		0.72

	That's X		mean		That's mean
N	0.26		0.04		0.30
V	0.01	+	0.51	=	0.52
A	0.73		0.45		**1.18**

Figure 1. Categorization of two instances of the ambiguous word "mean" using the co-clustering approach.

	Not X		cold		Not cold
N	0.29		0.23		0.52
V	0.04	+	0.05	=	0.09
A	0.67		0.72		**1.39**

	Not X		beans		Not beans
N	0.29		0.81		**1.10**
V	0.04	+	0.00	=	0.04
A	0.67		0.19		0.86

Figure 2. Categorization of two instances of the ambiguous template "Not X" using the co-clustering approach.

The categorization produced by co-clustering was evaluated in the same way as the one-dimensional clustering categorizations from the previous two experiments. The results for co-clustering with only full-utterance templates, and with both nested and full-utterance templates, are shown in the third and fourth column of Table 5 respectively. We can see that moving from one-dimensional clustering of templates to co-clustering of words and templates improves categorization correctness (as measured by F), both for the set of full-utterance templates and the extended set of full-utterance and nested templates. This result shows that making use of information from both the word and the context could allow a child to make a more accurate categorization of a word.

	Full-utterance templates (One-dimensional)	Full-utterance and nested templates (One-dimensional)	Full-utterance templates (Co-clustering)	***Full-utterance and nested templates (Co-clustering)***
Accuracy	0.787 (*0.457*)	0.798 (*0.497*)	0.857* (*0.457*)	0.832 (*0.497*)
Completeness	0.713 (*0.414*)	0.820 (*0.511*)	0.804 (*0.428*)	0.861 (*0.514*)
F	0.748 (*0.434*)	0.809 (*0.504*)	0.830 (*0.442*)	0.846 (*0.506*)

Table 5. Evaluation measures for lexical category assignment, showing both one-dimensional clustering and two-dimensional co-clustering, using data from full-utterance templates and from the extended set of full-utterance templates and nested templates. Baseline figures are shown in italics.

7. Concluding remarks

We have shown that the distributional co-occurrence of words with a set of lexically-specific templates, extracted in a straightforward manner from a corpus of child-directed speech, provides a computational model with enough information to induce word categories that strongly resemble the traditional lexical categories of nouns, verbs and adjectives.

A major challenge for theorists who suggest that children attend to the distributional contexts of words in order to induce lexical categories is to explain which contexts a child may plausibly attend to. The solution used by Mintz's (2003, 2006a, 2006b) frequent frames approach is to define a context "topologically", as the pair of words positionally flanking a target word. Our approach is to make use of structures that may be regarded as hypothetical linguistic constituents by the child, whether as full-utterance templates or as nested templates embedded in familiar contexts. These options are clearly not mutually exclusive, and we see no reason why children could not exploit both of these sources of information during language learning, and many others besides.

8. References

Brown, R. (1957). Linguistic determinism and the part of speech. *Journal of Abnormal and Social Psychology, 55*(1), 1-5.

Cameron-Faulkner, T., Lieven, E., & Tomasello, M. (2003). A construction-based analysis of child directed speech. *Cognitive Science, 27*, 843-873.

Goldberg, A. E. (1995). *Constructions: A Construction Grammar approach to argument structure*. Chicago: University of Chicago Press.

Gómez, R., & Maye, J. (2005). The developmental trajectory of nonadjacent dependency learning. *Infancy, 7*(2), 183-206.

Langacker, R. W. (1987). *Foundations of Cognitive Grammar, Vol. 1: Theoretical prerequisites*. Stanford, CA: Stanford University Press.

* The results shown for co-clustering differ slightly from the results presented in the poster version of this paper, as the results in the poster were obtained by using a set of automatic part-of-speech taggers for evaluation, rather than using the tags supplied with the Manchester corpus. Performance is roughly similar across the two sets of results.

Lieven, E., Pine, J. M., & Baldwin, G. (1997). Lexically-based learning and early grammatical development. *Journal of Child Language, 24*, 187-219.

MacWhinney, B. (2000). *The CHILDES Project: Tools for analyzing talk.* (3 ed. Vol. 2: The database). Mahwah, NJ: Lawrence Erlbaum.

Madeira, S. C., & Oliveira, A. L. (2004). Biclustering algorithms for biological data analysis: A survey. *IEEE Transactions on Computational Biology and Bioinformatics, 1*(1), 24-45.

Maratsos, M. P., & Chalkley, M. A. (1980). The internal language of children's syntax: The ontogenesis and representation of syntactic categories. In K. E. Nelson (Ed.), *Children's Language* (Vol. 2). New York: Gardner Press.

Mintz, T. H. (2003). Frequent frames as a cue for grammatical categories in child directed speech. *Cognition, 90*(1), 91-117.

Mintz, T. H. (2006a). Finding the verbs: Distributional cues to categories available to young learners. In K. Hirsh-Pasek & R. M. Golinkoff (Eds.), *Action meets word: How children learn verbs*. Oxford: Oxford University Press.

Mintz, T. H. (2006b). Frequent frames: Simple co-occurrence constructions and their links to linguistic structure. In E. V. Clark & B. F. Kelly (Eds.), *Constructions in acquisition*. Stanford: CSLI Publications.

Mintz, T. H., Newport, E. L., & Bever, T. G. (2002). The distributional structure of grammatical categories in speech to young children. *Cognitive Science, 26*, 393-424.

Nelson, K. (1995). The dual category problem in the acquisition of action words. In M. Tomasello & W. E. Merriman (Eds.), *Beyond Names for Things: Young Children's Acquisition of Verbs* (pp. 223-249). Hillsdale, NJ: Lawrence Erlbaum Associates.

Olguin, R., & Tomasello, M. (1993). Twenty-five-month-old children do not have a grammatical category of verb. *Cognitive Development, 8*, 245-272.

Pine, J. M., & Lieven, E. V. M. (1993). Reanalysing rote-learned phrases: Individual differences in the transition to multi-word speech. *Journal of Child Language, 20*, 551-571.

Pinker, S. (1979). Formal models of language learning. *Cognition, 7*, 217-283.

Redington, M., Chater, N., & Finch, S. (1998). Distributional information: A powerful cue for acquiring syntactic categories. *Cognitive Science, 22*(4), 425-469.

Shady, M. E. (1996). *Infants' sensitivity to function morphemes.* PhD thesis, State University of New York at Buffalo.

Sokal, R. R., & Sneath, P. H. A. (1963). *Principles of numerical taxonomy*. San Francisco: W. H. Freeman.

Theakston, A. L., Lieven, E., Pine, J. M., & Rowland, C. F. (2001). The role of performance limitations in the acquisition of verb-argument structure: an alternative account. *Journal of Child Language, 28*, 127-152.

Tomasello, M. (1992). *First verbs: A case study in early grammatical development*. Cambridge, UK: Cambridge University Press.

Tomasello, M. (2003). *Constructing a language: A usage-based theory of language acquisition*. Cambridge, MA: Harvard University Press.

Tomasello, M., & Olguin, R. (1993). Twenty-three-month-old children have a grammatical category of noun. *Cognitive Development, 8*, 451-464.

Van Zaanen, M. M. (2001). *Bootstrapping structure into language: Alignment-based learning.* PhD thesis, University of Leeds.